Opening the black box: Using explainable AI to understand what a neural network learns from lateral pinch simulations

Kalyn M. Kearney[1], Joel B. Harley[2], Jennifer A. Nichols[1]
[1]J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, United States
[2]Department of Electrical and Computer Engineering, University of Florida, Gainesville, United States
email: *kalynkearney@ufl.edu

Introduction

Machine learning approaches can infer complex biomechanical relations. Yet, many models are regarded as a "black box," lacking interpretability and limiting user confidence. While various works have recently advanced the field of explainable artificial intelligence (XAI)[1,2], very few have applied these methods to explain time-series biomechanical data. A notable example, Horst et. al[3] employed layer-wise relevance propagation to classify gait patterns from measured kinetic and kinematic data. This work presented a robust, interpretable model to provide data-driven gait analysis. XAI may provide clinical and scientific insights for a variety of other tasks beyond gait.

Here, we expand the use of XAI to the upper extremity, which is a complex, high degree-of-freedom system. Our objective was to reveal what a deep long short-term memory model (LSTM) learns from forward dynamic lateral pinch simulations. We used Shapley Additive Explanations (SHAP) to explain our LSTM's predictions. Known for consistent interpretations, SHAP considers all possible predictions for an observation using all possible combinations of features[4]. Importantly, we interpret SHAP values for our LSTM in the context of prior literature, providing both confidence in and insight from our model.

Methods

We developed an LSTM, which is a type of neural network, to predict lateral pinch thumb-tip forces from muscle activations. The LSTM included 14 inputs representing simulated muscle activations for 5 wrist and 9 thumb muscles. There were 3 hidden layers with 16 nodes each and 3 output nodes corresponding to 3-component thumb-tip forces. The LSTM used an RMSE loss criteria and an Adam optimizer. The LSTM underwent parameter tuning via random search, and 5-fold cross validation was used.

To provide observations to train our LSTM, we simulated lateral pinch data with varied anthropometric scaling and target forces. Using OpenSim v. 4.1, we scaled a thumb model[5] to random masses and bone lengths representing 5th-95th percentile young adults[6]. Each scaled thumb model was used in computed muscle control (CMC) simulations[7]. Target thumb-tip forces ranging from 40 N to 80 N in 5 N increments were CMC inputs. We applied these forces palmarly at the thumb-tip, as well as with 25% distal, 25% ulnar, and 25% radial deviations. Altogether, each scaled thumb model (525 total) was run through CMC 36 times (9 forces x 4 directions). Forward dynamics was then used to estimate thumb-tip forces resulting from the muscle activations from CMC. Muscle activations from CMC acted as LSTM inputs and thumb-tip forces from forward dynamics acted as outputs.

Data preprocessing included (1) removing simulations that failed to complete, (2) linearly interpolating all simulations to the same number of time points, (3) truncating the simulations to remove noise end effects, and (4) removing unphysical and unstable simulations. We then shuffled and split the resulting 6,590 simulations into training and testing datasets (80/20 split).

We analyzed the performance and predictions of the LSTM. Here, we report the RMSE of our LSTM evaluated on test data to elucidate the LSTM's ability to predict forces from muscle activations. To explain the predictions of the LSTM, we calculated SHAP values for 1000 random test observations. Briefly, SHAP values are calculated by first permuting all model features and training a distinct model (i.e. the LSTM) for each combination. The SHAP value for a feature is the average of the marginal contributions across all permutations of model features.

Results and Discussion

The LSTM predicted forces from muscle activations with low error. RMSEs for the LSTM were 1.64 N, 0.932 N, and 0.692 N for the distal, dorsal, and ulnar force directions, respectively. For each observation, the absolute error in the LSTM's prediction generally followed a normal distribution centered about 0 N.

The SHAP values of the five features most important for the LSTM's predictions are displayed in Fig. 1. Consistent with observations from cadaver specimens[8], the LSTM predicted a large negative dorsal force when activation of the *flexor pollicis longus* (FPL) was high and the *extensor pollicis brevis* (EPB) was low (Fig. 1, when FPL & EPB are red, SHAP value for F_y is large). Fascinatingly, the LSTM's prediction of thumb-tip force was substantially impacted by the activation of wrist muscles (Fig 1, *extensor carpi ulnaris*, ECU, and *flexor carpi ulnaris*, FCU). This result is consistent with prior literature, as wrist posture is known to affect lateral pinch strength[9].

In summary, we present two key findings: (1) the LSTM was able to learn the mapping between simulated muscle activations and simulated thumb-tip forces with low error and (2) interpreting the LSTM via SHAP revealed impacts of muscle activations on thumb-tip forces consistent with prior literature.

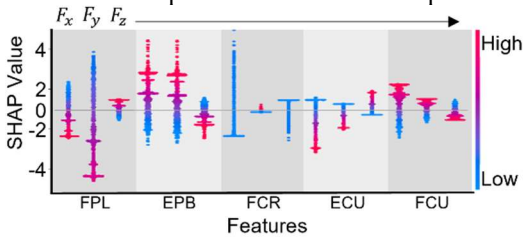

Fig. 1: SHAP values from 1000 test observations for the most important features predicting thumb-tip force in distal, dorsal, and ulnar directions (F_x, F_y, and F_z, respectively). Colors represent muscle activations.

Significance

The present work exemplifies not only the robustness of deep models for predicting upper extremity biomechanics, but that these models no longer need be considered a totally "black box."

Acknowledgments

Funding from National Science Foundation Graduate Research Fellowship and NIH NIBIB Trailblazer Award (R21EB030068).

References

[1] Roscher et al., 2020. *IEEE Access*. 8: 42200-42216. [2] Samek et al. 2017. *arXiv preprint*. [3] Horst et al. 2019. *Sci. Rep*. 9(1): 1-13. [4] Lundberg et al. 2017. *Proc. 31st NeurIPS*. [5] Nichols et al. 2017. *J Biomech*. 58: 97-104. [6] Fryar et al. 2016. *Vital Health Stat 3*. 39: 1-46. [7] Thelen et al. 2003. *J Biomech*. 36: 321-328. [8] Pearlman et al. 2004. *J Orthop Res*. 22(2): 306-312. [9] Dempsey et al. 1996. *Int. J. Ind. Ergon.* 17(3):273